DOCUMENT RESUME

ED 435 629 TM 030 227

AUTHOR Crehan, Kevin D.; Curfman, Mary
TITLE The Effect on Performance of Rapid Feedback on State Writing Assessments.
PUB DATE 1999-00-00
NOTE 21p.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Elementary School Teachers; English; *Feedback; Junior High Schools; Performance Based Assessment; Scores; *Scoring; *State Programs; *Test Results; Testing Programs; *Writing Tests
IDENTIFIERS *Nevada

ABSTRACT

The effect of rapid feedback for a state writing assessment on subsequent writing performance was investigated. In addition, the agreement between teachers' scores for the state writing assessment and state department scores was analyzed. Eighth grade English teachers (n=8) were trained in analytic trait scoring of writing assessments. They then scored their own students' state writing assessments soon after administration of the assessment. Teachers also scored assessments for a partner teacher's class. Normally, scores would not be available to teachers for several weeks following the assessment. A parallel writing assessment was administered to the classes of the eight participating teachers and to eight control classes. Results show good agreement between the teachers' scores and the scores assigned by the state department. There was fair agreement (76%) on the adequate-inadequate designation of student writing between the teachers and the state department. There was no difference between the writing performance for students of project teachers and students in the control classes. Teachers thought that the writing assessment was useful and that it would be more useful if results were received earlier in the school year. An appendix contains the teacher questionnaire. (Contains 3 tables and 18 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 629

Running head: EFFECT OF FEEDBACK ON WRITING ASSESSMENT

The Effect on Performance of Rapid Feedback On State Writing Assessments

Kevin D. Crehan

University of Nevada, Las Vegas

Mary Curfman

Clark County School District

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Kevin Crehan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030227

BEST COPY AVAILABLE

Abstract

The effect of rapid feedback for a state writing assessment on subsequent writing performance was investigated. Additionally, the agreement between teacher's scores for the state writing assessment and state department scores was compared. Eighth grade English teachers (N = 8) were trained in analytic trait scoring of writing assessments and scored their own students state writing assessments soon after administration of the state assessment. Normally scores on the assessment would not be available to teachers for several weeks following the assessment. Each teacher scored assessments for his/her class and the assessments for another partner teacher's class. A second parallel writing assessment was administered to the eight teacher's classes and eight additional control classes. Results showed good agreement between the teachers' scores and the scores assigned by the state department. There was fair agreement (76%) on adequate-inadequate designation of student writing between the teachers and state department. There was no difference between the writing performance for students of the project teachers and students in the control classes. Teachers felt the writing assessment was useful and would be more useful if results were received earlier in the school year.

The most prevalent response to the call for assessment reform has been an increase in the use of more authentic performance assessments. Advocates of performance assessments suggest that this from of appraisal can serve to measure important and complex learning outcomes and provide useful information to guide improvement in instruction (Resnick & Resnick, 1989). Perhaps the most complex form of student achievement that we attempt to assess involves composition. Therefore, the task of writing fits well within the realm of outcomes suitable for observation by performance assessments. Many states have added writing performance measures (among others) to supplement their assessment programs.

Among the problems associated with using performance assessments to measure important learning outcomes are objectivity of ratings and generalizability (reliability) of scores across raters and tasks. A review by Linn (1993) summarized evidence of acceptable generalizability across raters given well-defined scoring rubrics, intensive rater training, and monitoring during rating. Additionally, the California Assessment Program has established an interrater reliability of .90 for their writing assessment by using procedures which include providing sample anchor papers for each rater and recirclulating previously scored papers to check on stability (U.S. Congress, Office of Technology Assessment, 1992). Shavelson, Baxter, and Pine (1992) observed the reliability and validity of performance assessments in the 5th and 6th grade science curriculum. They asked the question: How large a sample of observers is needed to produce reliable measurement? Their results found

interrater reliability to be consistently high in evaluating student performance on complex tasks, high enough to conclude that a single rater provides a reliable score.

While these observations offer promise for the potential usefulness of performance assessments, scoring consistency is only one aspect of quality in decision situations based on assessment results. Linn and Burton (1994) suggest that for pass-fail decisions involving individual students, acceptable generalizability across tasks is attained only when a large number of tasks are used, perhaps as many as ten. If the content area being assessed is writing, such a large number of writing tasks on an occasion might require an unreasonable expenditure of instructional time devoted to assessment, to say nothing of the administration and scoring costs. However, increasing the number of ratings per task may yield an increase in "task" generalizability without a dramatic increase in the actual number of tasks. Multitrait analytic scoring strategies for writing performance assessments may increase "task" generalizability over a single, holistic, score.

Much of the research on the psychometric characteristics of writing performance assessments uses single score holistic ratings. In writing assessment this single holistic score is designed to estimate the wholeness in quality of the writing product. There is agreement (e.g., Huot, 1990) that writing is a multifaceted performance and, as such, involves attainment on a number of traits. Spandel and Stiggins (1994) suggest six traits on which the writing product differs: ideas, organization, voice, word choice, sentence fluency, and conventions. Additionally,

there are different types of writing, e.g., descriptive, persuasive, expository, narrative, and imaginative. Given that writing performance involves a number of types and traits on which individuals differ, some researchers (Roid, 1994; Huot, 1990; Marsh & Ireland, 1987; Novak, Herman, & Gearhart, 1996) recommend analytic scoring of writing products.

Roid (1994) used cluster analyses to explore the empirical validity of the previously named six analytic traits and found evidence of individual differences in trait profiles. Results of these analyses demonstrated that, while forty percent of the responses had flat trait patterns, a number of distinct patterns among the six traits were evidenced. For example, thirteen percent of the patterns were very close to average on five of the traits but either high or low on conventions. Ten percent of the patterns showed high or low voice, with other scores near average. An additional thirteen- percent was either high or low on ideas, organization, and voice but close to average on word choice, sentence fluency, and conventions; suggestive of a creative or stylistic component among the six traits. This evidence supports the potential usefulness of analytic scores as effective sources for feedback to students, guides to instruction, and as a basis for meaningful discussion of the writing process.

Work at the Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, at UCLA (e.g., Wolf & Gearhart, 1993a; 1993b) has expanded on the development of methodology and uses of analytic scoring. Work on narrative-writing-specific scoring rubrics has shown

promising evidence of reliability and validity (Gearhart, Herman, Novak, Wolf, & Abedi, 1994; Gearhart, Herman, & Novak, 1996). Additionally, training and use of these rubrics have benefited instruction by increasing participant teachers' understanding of the quality components of writing (Gearhart & Wolf, 1994; Gearhart, Wolf, Burkey, & Whittaker, 1994, Wolf & Gearhart, 1995).

Clearly, given the complexity of the writing task, the job of developing and implementing an analytically scored state writing assessment is enormous. However, if appropriately advanced, the outcomes of the process offer substantial benefit to the instructional process. Among the appropriate intentions of a state assessment program is teacher participation in the administration, development, and scoring of the assessments (Lane, Parke, and Stone, 1998). Unfortunately, while teacher involvement is present in state assessment programs, participation beyond administration is often limited to a relatively few teachers who score student writing.

Another appropriate feature of state assessment is rapid feedback to teachers to allow utilization of results for instructional decisions. Often the results of the assessment are received weeks or months after the administration, which limits the potential influence of the assessment results on instruction. Additionally, the limited involvement of teachers and the delay in receipt of results can have the consequence of the assessment being viewed as adjunct to instruction. The delay in receiving results takes on greater significance when the assessment has high stakes, such as a state department of education certification program for high school graduation. Since

this decision situation is so important, rapid feedback of assessment results to the classroom teacher is essential to allow the development and delivery of remediation to students at risk of failure. Given the significance of this decision and that writing is a part of the assessment, the direct involvement of teachers of composition in the process of establishing and implementing performance criteria is important. Additionally, it is essential that all composition teachers have training in the methods of analytic scoring and the utilization of analytic scores to improve instruction and provide feedback to students. The purpose of the present study is to assess the effects of teacher training in executing and using analytic scoring on the quality of their students' subsequent writing. More specifically the study addresses three questions:

1. How closely do the classroom teacher's scores agree with the state department of education scores?

If the classroom teacher's scores are not dramatically different than the state department scores, then the recommendation would be to modify the state scoring system to include the classroom teacher's score for his/her own students. This practice could reduce the costs of scoring.

2. Do students of teachers who are directly involved in the immediate scoring of their state writing assessment perform better on subsequent writing assessments?

If the results evidence a positive effect on student performance, it follows that the

recommendation would be to provide all teachers of composition with training on analytic scoring and the uses of the results in instruction.

3. Are teacher attitudes toward the value of the state assessment program affected by participation in this study?

In addition to the questions addressed directly in this study, another possible outcome is a shift in teacher perception of the state assessment program. Implementation of the program has accountability as a central theme and, as such, the value of the program to instruction is not always obvious to teachers. Perhaps closer involvement with the process will have an effect on teacher attitudes.

Methods

Subjects During the first week of school, an invitation to participate in the study went to all eighth grade English teachers. Eighth grade was chosen since this is a grade at which the state writing proficiency exam is given. The request was for two teachers of eighth grade English at the same school to work together on the project. Teachers were to be certified in English, not to have been previously trained by the state in analytic scoring, and to be currently teaching at least two sections of eighth grade English. Teachers were offered $150.00 for their efforts. Twelve teachers participated in the initial orientation and training. Two of these teachers subsequently withdrew from the study. As a result, ten teachers, located at five different middle schools participated in the study. Unfortunately, partial data for two of these teacher was lost and, as a result, complete data on eight teachers was available.

Training Teachers met on a Saturday in October and were given an orientation to the study and training in the analytic scoring model used by the state. This model uses four-traits: ideas, organization, voice, and conventions. Each trait is scored on a five-point scale. The orientation and training lasted approximately four hours including breaks.

Scoring At the close of the training session each teacher was given copies of the student papers from the state writing assessment conducted in late September for one of his/her classes and one of his/her partner's classes. Instructions were to score both sets of papers using the four trait analytic method. Following submission of their scores, teachers were encouraged to review their assessment results and discuss appropriate instructional approaches with other participating teachers and curriculum specialists.

During February a second writing assessment, which paralleled the previously administered state assessment, was administered to the same eight classes of students of participating teachers and to an additional group of eight control group classes. Each participating teacher scored his/her own student papers and their partner's student papers. Two participating teachers also scored each control group paper. At about the same time a brief questionnaire (see Appendix A) designed to assess teacher attitudes toward the statewide assessment program were administered to participating teachers and teachers of the control classes.

Results

The first question addresses the degree of agreement between the ratings of project teachers and the state department ratings. Two state department raters scored each student's writing product. If these scores were more than one point in difference, a third rater arbitrated so as to move the scores to within one point. The resulting two scores were then averaged to provide the reported score. Two project teachers also scored each paper and these scores were averaged without consideration of how widely the scores differed. The results of percent agreement between teacher scores and state department scores is reported in table 1. The

(Insert Table 1 about here)

degree of agreement in scores within .5 points is generally good at the extremes but falls off considerably for the middle ratings where the frequency of scores is greater. Agreement within 1.0 point is much better with all but a few of the percent agreements over ninety.

In addition to scores for each of the analytic traits, the state department reports whether the student writing was "adequate" or "inadequate." An adequate performance requires that all four-trait scores be at least three. If any of the four scores is below three, the student's writing is rated as inadequate. Table 2 presents

(Insert Table 2 about here)

the cross tabulation of these "pass-fail" decisions for both the state department scores and the teacher scores. There was a 76% agreement on this classification.

The second question addressed whether the students of the teachers involved

in this study would perform better on a subsequent (about four months later) writing task. Table 3 reports the descriptive statistics for the participating and control groups

(Insert Table 3 about here)

on the state assessment as scored by the state and the spring follow-up assessment as scored by the participating teachers. Analyses of covariance comparing the participating group to the control group on each of the four analytic traits for the spring assessment were conducted using the respective state assessment score as the covariate. None of the four tests was significant at the .05 level.

The last question concerned the possible impact on attitudes of participation in the study. Participating and control teachers at the end of the data collection completed a brief survey (see Appendix A). The responses were quite similar between the two groups of teachers with the predictable exceptions of statements 3 and 4, which addressed discussion of assessment results with students and colleagues. Since the participating teachers were encouraged to discuss the results, it's not surprising that they agreed with these statements. Overall, teachers felt the writing assessment was useful and would be more useful if the results were received earlier in the school year.

Discussion

The results of this small study offer no evidence of an effect on the quality of student writing associated with their teacher's involvement in the scoring of their state writing assessments early in the fall semester. Teacher comments on the survey

showed a desire to have the results sooner (results are currently received well into the spring semester) and to be trained in the analytic scoring method. Teachers also expressed a desire to receive training in the teaching of writing skills, especially in methods appropriate for at-risk students.

The results showed good agreement between teacher and state department ratings. Given only four hours of training and without the assistance of an arbitrator, the ratings of the participating teachers were very similar to the state department scores. Given these results, exploration of a scoring model that includes the classroom teacher is recommended. One such model would be to prepare one teacher "scoring leader" at each school location to coordinate training and scoring of the assessment at the school site. Scoring would occur very soon following administration of the assessment. The scores would then be forwarded to the state department and be immediately available to the classroom teacher (but not communicated to students or parents). The state department could then obtain an additional rating of the student writing product, compare this rating to that of the classroom teacher, and use an arbitrator to resolve discrepancies as is done currently. The savings in reduced scoring expense for the state department could be diverted to the school to reward the scoring leader and teachers involved in scoring. If successful, this model could be extended to other content areas in which state proficiency assessments are employed. As Lane, Parke, and Stone (1998) have suggested, teacher participation in the development, administration, and scoring of

assessments is an appropriate intention of a state assessment program.

References

Gearhart, M., & Wolf, S.A. (1994). Engaging teachers in assessment of their students' writing: The role of subject matter knowledge. Assessing Writing, 1, 67-90.

Gearhart, M., Herman, J. L., Novak, J. R., Wolf, S. A., & Abedi, J. (1994). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric (CSE Tech. Rep. 389). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J. L., & Novak, J. R. (1996). Issues in portfolio assessment: The scorability of narrative collections. (CSE Tech. Rep.) Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Wolf, S. A., Burkey, B., & Whittaker, A. K. (1994). Engaging teachers in assessment of their students' narrative writing: Impact on teachers' knowledge and practice (CSE Tech. Rep. 377). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. Review of Educational Research, 60, 237-263.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluation the consequences of assessment programs. Educational Measurement: Issues and Practice, 17, 24-28.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis, 15, 1-16.

Linn, R. L. & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13, 5-8, 15.

Marsh, H. W., & Ireland, R. (1987). The assessment of writing effectiveness: A multidimensional perspective. Australian Journal of Psychology, 39 (3), 353-367.

Novak, J. R., Herman, J. L., & Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of student writing. The Journal of Educational Research, 89 (4), 220-233.

Resnick, L. B., & Resnick, D. P. (1989). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M.C. O'Conner (Eds.), Future Assessments: Changing views of aptitude, achievement, and instruction (pp.37-75). Boston, MA: Kluwer.

Roid, G. H. (1994). Patterns of writing skills derived form cluster analysis of direct-writing assessments. Applied Measurement in Education, 7(2), 159-170.

Shavelson, R. J., Baxter, G. P., & Pine J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21, 22-27.

Spandel, V., & Stiggins, R. J. (1990). Creating writers: Linking assessment and writing instruction. White Plains, NY: Longman. instructional decisions.

U.S. Congress, Office of Technology Assessment. (1992) Testing in American Schools: Asking the Right Questions (OTA-SET-519), Washington, D.C.: U.S.

Government Printing Office.

Wolf, S. A., & Gearhart, M. (1993a). Writing What You Read: Assessment as a learning event (CSE Technical Report 358). Los Angeles: University of California. Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A., & Gearhart, M. (1993b). Writing What You Read: A guidebook for the assessment of children's narratives (CSE Resource Paper No. 10). Los Angeles: University of California. Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A., & Gearhart, M. (1995). Engaging teachers in assessment of their students' narrative writing: Patterns of impact and implications for professional development. Presented at the annual meeting of the American Educational Research Association, San Francisco.

Appendix A

CCSD/UNLV
ASSESSEMENT PROJECT

The following statements describe interest in and use of the Nevada State Writing Assessment. Please read each of the statements and respond using the scale below. Please respond candidly, and don't put your name on this form. Return completed form by March 10 to Mary Curfman, Secondary English Specialist, Curriculum & Professional Development, Secondary Education Division, North Ninth.

A. very much like me
B. like me
C. not like me
D. very much not like me

_____ 1. I don't use the assessment results in anyway.

_____ 2. I'm interested in seeing the results for my students.

_____ 3. I discuss the results with my colleagues.

_____ 4. I discuss the results with my students.

_____ 5. I don't use the results very much for instructional decisions because they arrive too late in the school year.

_____ 6. If the results were received early in the fall, I'd use them in planning writing instruction.

_____ 7. If the results were received early in the fall, I'd use them for individualized writing instructional programming.

_____ 8. I believe that the state assessment program, as structured, is useful for accountability purposes only.

Comments?

Table 1

Percent agreement of teacher ratings with state department (SDE) ratings within.5 points and 1.0 point for the four analytic traits: ideas (I), organization (O), voice (V), conventions (C).

SDE RATING	1	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
WITHIN .5 for I	100	100	63	69	73	83	87	100	100
WITHIN 1.0 for I	100	100	97	98	98	100	100	100	100
WITHIN .5 for O	100	80	62	62	73	79	94	80	75
WITHIN 1.0 for O	100	100	95	91	93	84	100	100	100
WITHIN .5 for V	---	67	62	56	76	90	60	100	75
WITHIN 1.0 for V	---	78	92	81	87	100	93	100	100
WITHIN .5 for C	80	70	58	86	76	55	94	73	100
WITHIN 1.0 for C	100	80	95	96	94	91	100	100	100

Table 2

Contingency table for percent "pass-fail" classification for state department (SDE) scoring and participating teacher scoring Sample size is 140.

		TEACHER SCORING	
		FAIL	PASS
SDE	FAIL	41	17
	PASS	8	34

Table 3

Means and standard deviations for the participating (P) and control (C) groups for the state writing assessment as scored by the state department (SDE) and the spring follow-up (SF-U) assessment scored by participating teachers on the four analytic traits: ideas (I), organization (O), voice (V), conventions (C). Sample sizes for P and C are 140 and 129 respectively.

		State Assessment			
		I	O	V	C
SDE	P	3.02	3.01	2.99	2.90
		(.78)	(.82)	(.78)	(.92)
SDE	C	3.30	3.19	3.19	3.15
		(.71)	(.72)	(.76)	(.82)
SF-U	P	3.39	3.40	3.53	3.29
		(.82)	(.95)	(.80)	(.95)
SF-U	C	3.41	3.26	3.50	3.08
		(.72)	(.81)	(.76)	(.85)